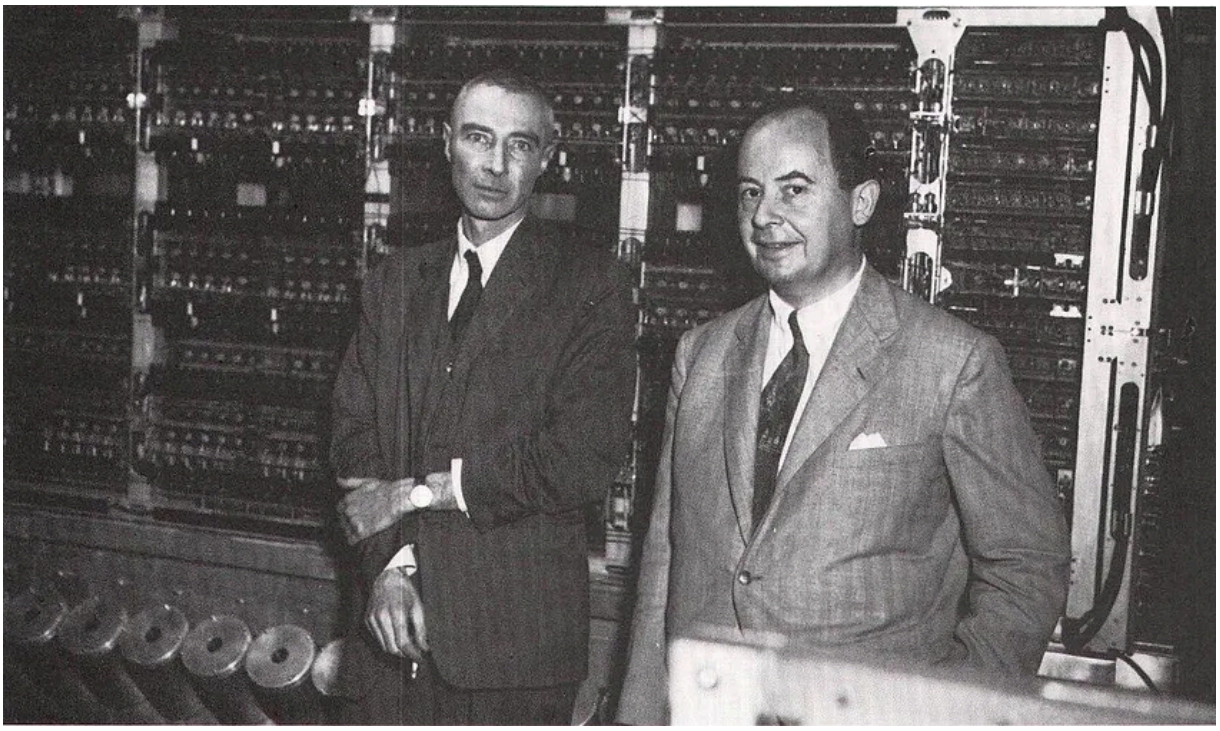
AI Isn’t Outthinking Mathematicians. It’s Out-Remembering Them.

davidepiffer.com/p/ai-isnt-outthinking-mathematicians

Davide Piffer August 4, 2026



At the 1952 dedication of the Institute for Advanced Study computer. AI may be less like an electronic Einstein than a machine-amplified von Neumann: immense speed, breadth and symbolic memory.

When an AI system solves a difficult mathematical problem, the usual explanation is that it has become more intelligent.

Perhaps it has absorbed millions of mathematical examples. Perhaps reinforcement learning has taught it better reasoning strategies. Perhaps it is beginning to develop something resembling genuine mathematical intuition.

All of these explanations may contain some truth. But they overlook a simpler possibility:

AI has access to a vastly larger working memory than the human brain.

Or, more precisely, it has access to an enormous external symbolic workspace that performs many of the functions that working memory performs in humans.

This difference may be especially important in mathematics.

A human mathematician can hold only a small number of unfamiliar elements in mind simultaneously. An AI model can keep the entire problem statement, hundreds of intermediate equations, several abandoned approaches, definitions, constraints and earlier conclusions inside its context window.

We normally interpret the resulting performance as evidence of superior reasoning. But some of it may instead reflect the removal of one of the most important biological limits on human reasoning: our extremely restricted working-memory capacity.

Mathematics is constrained by memory

Working memory is the mental system that allows us to hold and manipulate information over short periods.

When solving an equation, you must remember what each variable represents, which operations have already been performed and what the current goal is. During a proof, you may need to keep track of assumptions, intermediate lemmas, exceptions and multiple possible cases.

Human working memory is remarkably limited.

Its exact capacity depends on the task and on how information is organized, but the general limitation is obvious from everyday experience. Try multiplying two three-digit numbers in your head. The underlying operations are simple. The difficulty comes largely from having to preserve partial results while performing additional calculations.

Writing the numbers down transforms the problem.

Paper does not make you more intelligent. It expands your effective working memory.

The same principle applies at higher levels of mathematics. A mathematician uses notation, scratch paper, diagrams and previously written lemmas not merely to communicate the solution, but to make the reasoning cognitively possible.

Experts compensate through “chunking.” A novice sees a long sequence of symbols. An expert recognizes a familiar structure and treats it as a single conceptual object. This allows far more information to fit inside the same biological working-memory limit.

But chunking does not eliminate the limit. It merely compresses the information.

An AI model faces a very different constraint.

Working memory predicts mathematical performance beyond IQ

The importance of working memory for mathematics is not merely theoretical. It is visible in the differences between human beings.

Working memory is strongly related to general intelligence, which raises an obvious question: does it independently predict mathematical performance, or is it merely another imperfect measure of IQ?

Several studies suggest that it contributes something beyond conventional intelligence measures. Alloway and Passolunghi (2011), for example, examined working memory, verbal ability and mathematical skills in children. They found that working-memory measures made a distinct contribution to mathematical performance rather than simply reproducing the association between mathematics and general verbal ability.

In a separate six-year longitudinal study, Alloway and Alloway (2010) measured children at age five and then examined their academic achievement six years later. Early working-memory performance predicted later literacy and numeracy even after IQ was included in the analysis. Indeed, working memory was a stronger predictor of the later academic outcomes than the IQ measure used in the study.

Blankenship and colleagues (2015) similarly reported that working memory explained unique variation in mathematical fluency and calculation after statistically controlling for IQ and age. A large meta-analysis by Friso-van den Bos and colleagues (2013) also found a consistent relationship between working memory and mathematics across primary-school studies, although the strength of the relationship varied according to the type of working-memory and mathematical task being measured.

These findings should not be exaggerated. Working memory and intelligence overlap substantially, and statistical control cannot perfectly isolate them as independent psychological mechanisms. Nor does the evidence imply that commercially training working memory will necessarily produce large improvements in intelligence or mathematics.

The narrower conclusion is nevertheless important: **among children with similar measured intelligence, differences in the ability to hold, update and manipulate information still predict differences in mathematical performance.**

This provides a crucial clue for understanding AI. If human mathematical performance is partly capped by a working-memory bottleneck, then giving a machine an enormous symbolic workspace changes the nature of the contest. The machine may

appear more mathematically intelligent partly because it is much less constrained by a cognitive limitation that suppresses human performance.

The context window is a gigantic notebook

A modern language model can process an enormous sequence of tokens at once. This sequence may include the original question, definitions, examples, intermediate calculations and the model's own earlier reasoning.

The context window is not identical to human working memory. It is better understood as a gigantic external notebook combined with an imperfect system for searching and using what has been written in it.

This distinction matters.

Humans possess a form of active internal memory. We can silently choose a number, hold it in mind, transform it and replace it with a new value without saying or writing anything.

Standard language models are much weaker at maintaining this kind of private, continuously updated mental state. Their most stable form of memory is usually the sequence of tokens that has already been generated.

If the model writes:

$x=6$

and later writes:

$x+3=9$,

those statements remain inside the context. The model can attend to them again when generating the next step.

Its reasoning is therefore often externalized. The text is not merely a report of a completed thought process. The text is part of the mechanism by which the reasoning occurs.

Humans do something similar when using scratch paper. The major difference is scale.

An unaided human may struggle to keep five unfamiliar conditions active simultaneously. An AI can preserve dozens or hundreds of them in explicit form.

This does not mean that every item in a long context is retrieved perfectly. Models can overlook relevant information, become distracted or lose track of details. Advertised context length is not the same as perfectly usable memory.

Nevertheless, the difference in potential capacity is enormous.

Why this matters particularly for mathematics

The context-window advantage is not equally useful in every kind of reasoning.

It matters especially for mathematics because mathematical reasoning can be translated unusually well into explicit symbols.

Almost every relevant element of a mathematical problem can be written down:

- the assumptions;
- the definitions;
- the known equations;
- the current objective;
- the results already proved;
- the cases that have been eliminated;
- the conditions under which each step remains valid.

Once written, this information remains stable.

If (x) is defined as an integer at the beginning of a proof, it remains an integer unless the proof explicitly changes the definition. A strict inequality does not gradually become a non-strict inequality because of changes in mood, context or interpretation.

Mathematical symbols are designed to reduce ambiguity.

This makes mathematics almost perfectly suited to an intelligence that operates through a large textual workspace.

Consider a problem requiring the solver to remember that:

- n is odd;
- p is prime;
- $x \neq 0$

and that one branch of the argument has already produced a contradiction.

A human may understand the basic strategy but divide by x before establishing that $x \neq 0$. The error is not necessarily caused by a lack of intelligence. It may be a failure of bookkeeping.

An AI can restate the active constraints at each stage:

> We are working under the assumptions that (n) is odd, (p) is prime and ($x \neq 0$).

The context becomes a ledger of the reasoning state.

Many difficult mathematical problems contain a profound insight somewhere near the beginning, but they also contain a large amount of less glamorous work afterward: expanding expressions, checking cases, carrying conditions through transformations and making sure that the final conclusion is compatible with every earlier assumption.

A machine does not need to possess deeper insight than a human to gain an advantage here. It may simply be better equipped to preserve the entire state of the problem while completing a long sequence of operations.

Long reasoning chains can exceed human capacity

Mathematics is highly compositional.

A proof can often be represented as:

$A \rightarrow B \rightarrow C \rightarrow D$

If each step is valid and the chain is preserved accurately, the conclusion follows.

A large working space allows the model to construct much longer chains before losing the thread.

This is important because the difficulty of a problem does not depend only on the difficulty of each individual step. It also depends on how many steps must be coordinated.

A person may be perfectly capable of understanding every local inference in a 100-step argument while still being unable to generate the entire argument unaided. The problem exceeds the person’s ability to maintain the global structure.

AI can potentially compensate by writing down nearly everything.

This may explain why additional “thinking time” often improves model performance. More computation allows the system to produce more intermediate states, examine

alternative branches and preserve partial conclusions.

What looks like deeper thought may sometimes be broader search conducted inside a much larger notebook.

Informal reasoning does not work the same way

Now compare a mathematical problem with a social question:

> Why has Maria suddenly stopped replying to my messages?

A larger context window might allow an AI to examine years of correspondence. It could identify changes in tone, timing and vocabulary.

But the decisive information may still be missing.

Perhaps Maria is angry. Perhaps she is busy. Perhaps she is ill. Perhaps she has lost her phone. Perhaps she is avoiding an unrelated problem.

No amount of memory can retrieve facts that were never observed.

The challenge is not simply to preserve a known set of premises and derive their consequences. It is to reason under uncertainty about hidden causes.

Informal reasoning also depends heavily on concepts whose meanings are unstable.

Words such as “fair,” “successful,” “responsible,” “harmful” or “intelligent” do not possess the exactness of mathematical variables. Their meaning depends on culture, goals and context.

A model can remember every sentence in a discussion while still misunderstanding what the participants mean.

The same applies to political analysis, historical interpretation, business strategy and psychological judgment. In these domains, the central problem is often not working-memory capacity. It is identifying the correct causal model when the evidence is incomplete and ambiguous.

A larger notebook helps, but it does not solve the fundamental problem.

Mathematics is different because the reasoning environment is artificially constructed to make assumptions explicit and transformations verifiable.

Mathematics provides unusually strong feedback

Mathematical answers can often be checked.

A solution to an equation can be substituted back into the original expression. A proposed identity can be evaluated numerically. A computer program can test hundreds of cases. A formal proof assistant can verify whether every inference follows from accepted rules.

This creates an ideal environment for AI.

The model can use its context not only to produce a solution but also to record alternative approaches, identify contradictions and correct previous mistakes.

In less formal domains, feedback is much weaker.

A persuasive historical explanation may be impossible to verify conclusively. A business strategy may not reveal whether it was correct until years later. A psychological interpretation may remain permanently uncertain.

In such fields, a long and coherent argument can still be completely wrong.

Mathematics rewards systems that can generate, store and verify explicit intermediate states. These are precisely the abilities that context windows, scratchpads, code execution and formal verification amplify.

Is this really working memory?

Some researchers would object to describing a context window as working memory.

They have a point.

Human working memory is not merely a storage buffer. It involves attention, inhibition, continuous updating and the active manipulation of internal representations.

A language model's context is more passive. Previous tokens remain fixed. The model cannot literally return to an earlier line and replace it. It generates new tokens conditioned on the existing record.

The most accurate description may therefore be **augmented symbolic working memory**.

The model has weaker private mental memory than a human in some respects, but vastly greater explicit memory in others.

Humans are better at silently maintaining a small internal state. AI is better at operating over an enormous written record.

For mathematics, the second ability may often be more valuable.

Formal reasoning does not require every relevant state to remain private. On the contrary, mathematics improves when assumptions and intermediate steps are made explicit.

The apparent weakness of AI—its tendency to “think out loud”—becomes an advantage when the domain itself is built from written symbols.

Intelligence or cognitive architecture?

This leads to a broader question: what does it mean to say that AI is “better at mathematics” than humans?

Imagine giving two people the same mathematical problem.

One must solve it entirely in their head. The other receives unlimited paper, perfect notes, instant access to every previous calculation, the ability to try several approaches in parallel and a machine that checks each result.

If the second person wins, we would not necessarily conclude that the second person possesses greater raw mathematical intelligence. We might instead conclude that the second person had a much better cognitive architecture for the task.

The same caution should apply when comparing humans and AI.

AI systems are trained on enormous quantities of mathematical material. They can generate many possible solutions, use code, consult tools and preserve long reasoning traces.

Their performance is therefore the product of several factors: mathematical knowledge, reasoning ability, memory capacity, search, and verification.

We tend to focus on reasoning because it is the most philosophically exciting component. But memory may be doing much more of the work than we realize.

A testable prediction

The working-memory hypothesis produces clear predictions.

AI’s advantage should be largest on problems that involve:

- many interacting constraints;
- long calculations;
- extensive case analysis;

- repeated reference to previous results;
- exact symbolic bookkeeping;
- large bodies of formal definitions.

Its advantage should be smaller on problems that depend primarily on one short conceptual leap.

A brilliant mathematician may still outperform AI when the crucial challenge is finding an entirely new representation of a problem. Once that representation has been found, however, AI may be superior at exploring all its consequences.

The hypothesis also predicts that reducing a model's usable context or preventing it from writing intermediate steps should disproportionately damage performance on long mathematical tasks.

Conversely, expanding a human's external memory—with clear notation, diagrams, software and structured notes—should narrow part of the gap.

The fairest comparison may therefore not be AI against a human thinking unaided. It may be AI with its tools against a human with equally powerful external memory and verification systems.

AI may be out-remembering us

The rise of mathematical AI is often described as a triumph of machine intelligence over human intelligence.

That interpretation may be premature.

AI certainly appears to be learning better reasoning strategies. It may eventually develop forms of mathematical intuition that equal or exceed our own.

But some of its present advantage may be more prosaic.

The human brain evolved under severe constraints. It has limited working memory, becomes tired, loses intermediate results and struggles to coordinate long chains of unfamiliar symbols.

AI is not bound by the same architecture.

It can turn reasoning into text and use that text as a vast external cognitive workspace. Mathematics, because of its precision and symbolic structure, is the domain where this advantage is most easily converted into superior performance.

The real reason AI is beating humans at mathematics may therefore be less mysterious than we imagine.

Perhaps the most revealing comparison is not between AI and the average human, but between two very different forms of genius.

The physicist Eugene Wigner, who knew both John von Neumann and Albert Einstein, wrote that no one he had encountered possessed a mind as “quick and acute” as von Neumann’s. Von Neumann could absorb vast amounts of information, follow extraordinarily complicated arguments and move between mathematical fields with astonishing speed. Yet Wigner still regarded Einstein’s understanding as deeper, more penetrating and more original. Von Neumann may have had the greater raw intellectual processing capacity, but Einstein was more likely to reconceptualize the problem itself.

Present-day AI appears more like a machine-amplified version of the first set of abilities than the second. It is extraordinarily fast, broad and capable of preserving and manipulating huge quantities of explicit information. It can search many possibilities, remember long chains of deductions and execute formal reasoning at a scale no unaided human can match. But that does not necessarily mean it possesses Einsteinian depth: the ability to discard the accepted framework, identify the hidden conceptual error and invent a radically new way of seeing reality.

The analogy should not be pushed too far. Von Neumann was himself an immensely original thinker, not merely a fast calculator. But Wigner’s distinction captures something important. AI may currently be beating humans at mathematics less by thinking like Einstein than by combining something resembling von Neumann’s speed and breadth with a practically unlimited notebook.

The next great threshold will be reached when AI can do more than solve existing problems faster than humans. It will be reached when AI can recognize that a problem has been framed incorrectly and invent a fundamentally better way of understanding it.

References

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20–29. DOI: 10.1016/j.jecp.2009.11.003.

Alloway, T. P., & Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences, 21*(1), 133–137. DOI: 10.1016/j.lindif.2010.09.013.

Blankenship, T. L., O'Neill, M., Ross, A., & Bell, M. A. (2015). Working memory and recollection contribute to academic achievement. *Learning and Individual Differences, 43*, 164–169. DOI: 10.1016/j.lindif.2015.08.020.

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10*, 29–44. DOI: 10.1016/j.edurev.2013.05.003.